

# DynaMod: dynamic functional modularity analysis

Choong-Hyun Sun<sup>1</sup>, Taeho Hwang<sup>2</sup>, Kimin Oh<sup>3</sup> and Gwan-Su Yi<sup>1,2,3,\*</sup>

<sup>1</sup>Department of Computer Science, <sup>2</sup>Department of Information and Communications Engineering and

<sup>3</sup>Department of Bio and Brain Engineering, KAIST, Daejeon 305-701, South Korea

Received March 1, 2010; Revised April 17, 2010; Accepted April 24, 2010

## ABSTRACT

A comprehensive analysis of enriched functional categories in differentially expressed genes is important to extract the underlying biological processes of genome-wide expression profiles. Moreover, identification of the network of significant functional modules in these dynamic processes is an interesting challenge. This study introduces DynaMod, a web-based application that identifies significant functional modules reflecting the change of modularity and differential expressions that are correlated with gene expression profiles under different conditions. DynaMod allows the inspection of a wide variety of functional modules such as the biological pathways, transcriptional factor-target gene groups, microRNA-target gene groups, protein complexes and hub networks involved in protein interactome. The statistical significance of dynamic functional modularity is scored based on Z-statistics from the average of mutual information (MI) changes of involved gene pairs under different conditions. Significantly correlated gene pairs among the functional modules are used to generate a correlated network of functional categories. In addition to these main goals, this scoring strategy supports better performance to detect significant genes in microarray analyses, as the scores of correlated genes show the superior characteristics of the significance analysis compared with those of individual genes. DynaMod also offers cross-comparison between different analysis outputs. DynaMod is freely accessible at <http://piech.kaist.ac.kr/dynamod>.

## INTRODUCTION

The analysis of genome-wide gene expression profiles is widely used in the current biomedical research. An important goal of this analysis is to generate biological

hypotheses by identifying statistically significant genes and their functions that reflect the different phenotype classes under test. Various useful gene selection tools are available for this purpose (1,2). Typical gene selection approaches focus on the characterization of individual genes distinguishing the biological conditions. These individually selected genes are subjected to annotate enriched functional categories. However, many phenotypes and underlying cellular processes are associated with groups of genes in various functional categories and their networks rather than with isolated genes. Under the typical gene selection procedure, the coordinated effect of correlated genes in the functional modules or networks cannot be well characterized. Furthermore, the statistical significance of individual genes cannot be presented properly when the sets of genes in the test have expression dependencies that may exist in the functionally correlated genes. In fact, statistical test showed that the average significances of functionally correlated genes are closer to a normal distribution than those of individual genes (3). The coordinated changes of functionally correlated gene sets appears to be an improved measure of extracting significant causes from differential gene expression profile in the sense of biological and statistical significance at the same time.

Several gene set enrichment analyses (GSEAs) have been introduced (3–8). These tools emphasize the discovery of functionally correlated genes more reliably and completely than individual gene-based approaches. The typical strategy of the gene set-based approach is to combine the initial significance evaluation of all genes to the coordinated significance of the set of genes obtained from predefined functional categories. Although GSEA approach can more appropriately determine the potent effects of functionally correlated genes than individual gene-based approach, the efficiency of this method can vary depending on the strategy of selecting and scoring the gene sets. Most scoring methods used in GSEA tools remain limited, in that they must use a relatively large size of gene set in conjunction with a parametric approach (3,5). Alternative non-parametric approaches are usually associated with an increase in the computational

\*To whom correspondence should be addressed. Tel: +82 42 350 6160; Fax: +82 42 350 6814; Email: gsyi@kaist.ac.kr

complexity and a decrease in the level of the generality in comparisons of results from different tests. Another restriction to use current GSEA approaches is on the coverage of functional categories they choose. More importantly, they do not handle the networked characteristic of functional modules. A networked feature of the genes in a module can be characterized by the dynamic changes of correlated activities or biologically dysregulated relationships of the genes under different conditions (9,10). This can be applied to various functional modules to identify dynamic functional modularity (DFM), which could not be described in typical GSEA approaches. Another networked characteristic can be obtained by hub protein networks or protein complexes that connect the functional module sets of the analysis.

DynaMod is a web-based application that identifies significant functional modules reflecting the dynamic changes of correlated activities and differential expressions in gene expression profiles under different conditions. A novel scoring strategy uses the average mutual information (MI) differences of gene pairs in predefined functional modules depending on phenotypes. The gene sets are constructed from a wide variety of functional categories representing cellular pathways, transcription factor regulations, miRNA regulations, hub protein networks and protein complexes. Overlap genes across the functional modules are used to find interconnectivity across functional modules. It also provides a cross-comparison between different analysis outputs that helps users to interpret the results from more than two phenotypes or different functional categories together.

## METHODS AND IMPLEMENTATION

### Collection of functional modules and biological gene pairs

There are five types of functional modules in this system: biological pathways from KEGG (11), transcriptional factor–target gene groups from TRANSFAC (12), microRNA–target gene groups from MsigDB (4), protein complexes and hub networks from protein interactome. Integrated protein complexes from COFECO (13) and combined protein interaction database (14) are used here. Protein complexes are increased by adding human protein complexes from Ewing *et al.* (15). Hub networks with at least five interacting partners are retrieved from an integrated interactome including BIND (16), BioGrid (17), DIP (18), HPRD (19), IntAct (20), MINT (21) and MIPS MPPI (22). Gene pairs are constructed from the pairwise relations presented in these five types of functional modules and are utilized to compute MI. DynaMod imports flat files related to functional modules and their gene pairs. In addition, three disease databases including OMIM (23), Genetic Association Database (24) and Cancer Gene Census (25) are used to annotate disease genes in functional modules. All gene entries for the functional modules of DynaMod are marked with an Entrez Gene ID. Gene identifiers from various databases including Entrez Gene, UniGene, RefSeq, EMBL, ENSEMBL, SGD, RGD, MGI, HGNC and the microarray probe identifiers of Affymetrix and

Agilent are allowed. DynaMod also accepts identifiers of UniProtKB, iProClass and IPI. The same Entrez Gene is frequently represented by several probes in a microarray data set. DynaMod determines a probe representing a gene with the best scored gene probe when there are several probes corresponding to a gene in a microarray data set.

### Significance test of DFM

Our proposed DFM analysis works in two steps. We first use an aforementioned compendium of gene pairs to compute MI. MI is a quantity that measures the mutual dependence of two variables in information theory, which is zero if and only if the two variables are independent. The MI of gene pairs shows a correlation or dependence between two genes in the phenotype of interest. We then identify DFM of predefined modules according to the statistical significance of the altered correlation of gene pairs in the functional modules. To perform the latter step, MI difference ( $\Delta MI$ ) is measured by a method similar to that of Mani *et al.* (9), which proposed an oncogene prediction method using dysregulated interactions that show significant MI differences in the phenotype of interest. MI difference is given by the following equation:

$$\Delta MI = MI_{\text{total}} - MI_{\text{control}}$$

Here,  $MI_{\text{total}}$  is the MI calculated from all given samples and  $MI_{\text{control}}$  is the MI calculated from control sample set that excludes the phenotype samples of interest. The MI differences of gene pairs represent the change of correlated activity or biologically dysregulated relationships among the genes and the dynamic modularity can be measured by the average MI differences of functional modules. It is assumed that given an expression profile, there are  $N$  predefined gene pairs in each module type. Subsequently,  $N$  MI differences are computed. A null distribution is generated by sampling a subset of gene pairs across 100 equally sized MI difference bins covering  $N$  MI difference range in overall gene pairs of an expression profile. For each bin of 100 gene pairs, MI differences for those gene pairs are computed by phenotype randomization of experimental samples. Thereafter, a null distribution composed of resulting 10 000 MI differences is constructed. A normality test was conducted by several standard routines on four microarray data sets. The null distribution of MI differences for individual gene pairs in glioblastoma data set [GSE4290 from Gene Expression Omnibus (GEO) in NCBI] was slightly out of normal distribution according to Kolmogorov–Smirnov normality test ( $D = 0.0094$ ,  $P = 0.04493$ ), although it was much closer to normal than the case using other previously used measure such as fold change ( $D = 0.0819$ ,  $P = 2.2e-16$ ). After the test of increasing number of gene pairs, the sufficient minimal size of the gene pairs was set to two, which shows confident normality ( $D = 0.0082$ ,  $P = 0.2036$  according to Kolmogorov–Smirnov normality test). This result is comparable with the normality characteristic of the previous method acquired from the mean of 10 samples with a fold change measure (Supplementary Data). DynaMod identifies significant functional

modules in different conditions by using the  $Z$ -statistics of the average MI difference. This tool computes a  $Z$ -score for the average MI difference of a functional module using the aforementioned null distribution of MI differences and estimates the statistical significance of the  $Z$ -score against a standard normal distribution. If a sample size of MI differences is  $k$  (i.e.  $\geq 2$ ), the mean of the averages of all MI differences ( $\mu$ ) and the standard deviation of the averages of all of MI differences ( $\sigma$ ) are computed from a null distribution. When the mean of the MI differences for a given functional module is  $\overline{\Delta MI}$  and the number of gene pairs in that module is  $n$ , the standard error (SE) of  $\overline{\Delta MI}$  and the  $Z$ -score is computed as follows:

$$SE = \frac{\sigma}{\sqrt{n}}, \quad Z = \frac{\overline{\Delta MI} - \mu}{SE},$$

The statistical significances ( $P$ -values) of the functional modules are adjusted for multiple-testing routines, such as with the Bonferroni method or the false discovery rate (FDR) of Benjamini and Hochberg (26).

#### Determination of module expression activity

Module expression activity (MEA) is a score estimating the differential expression of a module and upregulated group (Up\_MEA) or downregulated group (Down\_MEA) of a module. MEA is given by the following equations.

$$\begin{aligned} Up\_MEA &= \frac{\sum_{i=1}^n g_i \times (\text{sgn}(g_i)+1)}{2n}, \\ Down\_MEA &= \frac{\sum_{i=1}^n g_i \times (1 - \text{sgn}(g_i))}{2n} \end{aligned}$$

Here,  $g_i$  is a gene score acquired by  $t$ -test or fold change, sgn is an indicator function and  $n$  is the number of member genes in the module. In Up\_MEA, sgn results in 1 if  $g_i$  is positive and 0 otherwise. In Down\_MEA, sgn results in -1 if  $g_i$  is negative and 0 otherwise.

#### Network study among significant functional modules

Unions of the functional and neighbor modules are provided for the significant functional modules, whose overlap score exceeds a specific threshold (one overlap gene as the default number). Alternative score is defined as  $(n \times n)/n_1 \times n_2$ , where  $n$ ,  $n_1$ , and  $n_2$  are the number of genes in the overlap and modules 1 and 2, respectively. Users can acquire the association among functional modules by genes in the overlap.

#### Significance test for individual genes and gene pairs

Although functional module-based analysis is the main goal, users may need occasionally the significance of individual genes. DynaMod evaluates the significance scores of individual genes by  $t$ -test or fold changes. DynaMod also provides the significances of  $\Delta MI$  for individual gene pairs.

#### Implementation

The core algorithm of DynaMod was implemented in R and the web interface was implemented in JAVA and Java Server Page on Linux. It runs on Apache Web Server combined with the Tomcat servlet engine. All annotation data for gene entries and functional modules within this system were stored in Oracle 10g RDBMS. As the computation works of DynaMod runs on a Linux-based cluster system, a large number of MI calculations of gene pairs can be parallelized using the Parallel Virtual Machine (PVM) via the rpvm and snow in R packages on a cluster of nine nodes, each with dual quad-core Intel Xeon 2.46 GHz CPUs and 24 GB of RAM. The functional modules, biological gene pairs, annotation resources, organisms and identifiers will be updated periodically.

#### DYNAMOD WEB SERVER

##### Input

The input of DynaMod is a genome-wide expression profile. Expression profiles have to contain gene identifiers, expression values and class labels (i.e. 1 or 2) of experimental samples.

##### Outputs

The following DynaMod outputs are accessible at specified URL addresses that are sent to user by e-mail. These outputs are also downloadable from the web pages. An example outputs are shown in Figures 1 and 2.

*Summary of significant functional modules, their networks and involved genes.* DynaMod produces a summary rank table of significant functional modules and their neighbor modules including their significance scores and Up/Down\_MEA scores. The summary table is linked to the further detail tables that contain the information of entry genes, gene pairs and their scores with  $P$ -values in the functional modules and networked modules. Biological functions of significant modules can be efficiently annotated by composite functional enrichment of COFECO (13).

*Graphical representation of functional modules.* Individual functional modules and their neighbor modules are summarized with a network graphical view of the scored genes and gene pairs. Up/down expression of genes and up/down expression correlations are depicted from green to red. The network graph is implemented using the GraphViz library.

*Cross-comparison between different analysis results.* A cross-comparison is used to identify changes/trends between the analysis results from different phenotypes. This functionality is useful for comparing various types of outputs from different input sets.

A	Rank	Module (exp_genes/tot_genes)	DFMA		MEA	Up_MEA			Down_MEA			Neighbor modules
			z_score	P_val	P_val	Genes	Activity	P-val	Genes	Activity	P_val	
1	PA700-20S-PA28_complex(36/36)	-10.206	0.0	1.164e-05	29	3.611	1.876e-13	7	-0.793	7.770e-07		
2	MCM_complex(6/6)	8.068	1.703e-14	2.714e-18	6	8.663	6.602e-19	0	0	1.0		
3	P2X7_receptor_signalling_complex(12/12)	7.843	1.059e-13	0.6933	7	2.197	6.277e-10	5	-1.729	5.242e-07		
4	40S_ribosomal_subunit(33/34)	-7.565	8.923e-13	5.328e-11	29	4.519	7.477e-13	4	-0.176	0.0156		
5	LSD1_complex(14/14)	7.216	1.161e-11	2.557e-07	10	3.175	9.626e-15	4	-0.622	1.0E-4		
6	PA700(20/20)	-7.022	4.236e-11	0.0757	15	2.122	6.494e-08	5	-1.05	9.242e-07		
7	28S_ribosomal_subunit(30/30)	-7.005	4.641e-11	2.026e-06	24	3.021	6.356e-17	6	-0.704	5.290e-06		
8	SNARE_complex_(VAMP2(4/4)	6.795	1.942e-10	9.382e-09	0	0	1.0	4	-6.429	2.039e-08		

  

B	Rank	Module (exp_genes/tot_genes)	DFMA		MEA	Up_MEA			Down_MEA			Neighbor modules
			z_score	P_val	P_val	Genes	Activity	P-val	Genes	Activity	P_val	
1	Systemic_lupus_erythematosus(131/140)	14.607	0.0	1.929e-20	112	3.857	2.901e-25	19	-0.431	1.257e-21		
2	Spliceosome(126/127)	-13.811	0.0	3.058e-13	91	3.815	3.479e-24	35	-0.781	4.456e-09		
3	Ribosome(95/98)	-28.022	0.0	1.072e-12	80	4.498	1.141e-16	15	-0.37	1.117e-09		
4	Viral_myocarditis(74/74)	28.371	0.0	3.108e-10	53	3.46	9.063e-18	21	-0.852	4.403e-15		
5	Proteasome(43/44)	-11.92	0.0	5.410e-09	35	4.253	6.005e-18	8	-0.646	2.449e-07		

  

C	Rank	Module (exp_genes/tot_genes)	DFMA		MEA	Up_MEA			Down_MEA			Neighbor modules
			z_score	P_val	P_val	Genes	Activity	P-val	Genes	Activity	P_val	
1	MIR27B(492/496)	-8.744	0.0	0.0543	267	2.272	2.973e-33	225	-1.931	5.381e-26		
2	MIR27A(492/496)	-8.744	0.0	0.0543	267	2.272	2.973e-33	225	-1.931	5.381e-26		
3	MIR524(463/470)	-9.092	0.0	0.313	255	2.215	2.292e-35	208	-1.77	1.836e-24		
4	MIR182(381/386)	-7.597	6.100e-13	0.0026	194	1.994	1.870e-33	187	-1.962	2.952e-18		
5	MIR1(308/309)	-6.534	1.075e-09	0.0222	159	2.171	2.869e-36	149	-1.888	4.121e-21		

**Figure 1.** Screenshots of the selected DynaMod analysis results with the example of glioblastoma (GSE4290 of GEO) data set. Highly significant modules in DFM are listed with their modular expression activity showing modular differential expression activity of a module (MEA) and upregulated group (Up\_MEA) or downregulated group (Down\_MEA) of a module for the functional categories of protein complex (**A**), KEGG pathway (**B**) and miRNA targets (**C**). The name of module in ‘Module’ column links the detail information of each module as shown in Figure 2A–C. The icon in ‘Neighbor modules’ column links the detail neighbor information as shown in Figure 2D.

## AN EXAMPLE APPLICATION

Figures 1 and 2 show the selected DynaMod outputs for an example analysis of glioblastoma data set (GSE4290) that have been submitted to the GEO database at NCBI (27). The functional modularity of PA700-20S-PA28 complex, one of the proteasome complexes, looks decreased mostly [most negative Z-score of MI difference in DFM activity (DFMA) column] and the member genes are separated into up- and downregulated groups during the change from normal to glioblastoma. For the second ranked MCM complex, it increases mostly (most positive Z-score) and all member genes are upregulated during the

cancer formation. These results can give new implications of the role of those complexes in conjunction with previous studies indicating that PA700-20S-PA28 complex is involved in tumorigenesis and immune surveillance (28) and MCM complex is highly expressed in malignant human cancer cells (29). The eighth ranked complex, ‘SNARE complex’, is sublocalized in neurons of brain and is responsible for membrane fusion in the secretory pathway (30). All member genes are downregulated with significant decrease of their modularity in glioblastoma. Detail information of SNARE complex is shown in Figure 2. With the inspection of

A	Rank	ProbeID(or symbol)	Official symbol	Description	Score	p_value	Corr_pval
1	204729_s_at	STX1A GAD		syntaxin 1A (brain)	-10.367	3.6860e-11	4.8746e-10
2	1556629_a_at	SNAP25 GAD		synaptosomal-associated protein, 25kDa	-6.984	3.7143e-08	2.8080e-07
3	201556_s_at	VAMP2 GAD		vesicle-associated membrane protein 2 (synaptobrev...	-4.673	6.8550e-05	2.6E-4
4	206368_at	CPLX2 GAD		complexin 2	-3.693	0.0011	0.00259

  

B	Rank	Gene pairs	Type	MI difference	p_value	Corr_p_val
1		204729_s_at--201556_s_at (STX1A--VAMP2) GAD	PPI, co-complex	0.179	4.818e-05	0.1337
2		206368_at--201556_s_at (CPLX2--VAMP2) GAD	co-complex	0.115	3.0E-4	0.2177
3		206368_at--204729_s_at (CPLX2--STX1A) GAD	PPI, co-complex	0.057	0.0013	0.2995
4		206368_at--1556629_a_at (CPLX2--SNAP25) GAD	co-complex	0.017	0.0031	0.3413
5		1556629_a_at--204729_s_at (SNAP25--STX1A) GAD	PPI, co-complex	-0.118	0.039	0.4742
6		1556629_a_at--201556_s_at (SNAP25--VAMP2) GAD	PPI, co-complex	-0.158	0.0701	0.4858

  

C	D																																																							
<p>Functional module network: Down  Up  Not expressed </p> <p>SNARE_complex_(VAMP2,_SNAP25,_STX1a,_CPLX2)(cc_794)</p>	<table border="1"> <thead> <tr> <th>Neighbor modules</th> <th>Type</th> <th>Module size</th> <th>Overlap size</th> <th>Overlap genes</th> </tr> </thead> <tbody> <tr> <td>STX1A</td> <td>hub</td> <td>59.0</td> <td>4.0</td> <td>CPLX2, SNAP25, STX1A, VAMP2</td> </tr> <tr> <td>MIR1</td> <td>miRNA</td> <td>309.0</td> <td>3.0</td> <td>CPLX2, SNAP25, VAMP2</td> </tr> <tr> <td>SNARE_complex_(VAMP2,_SNAP25,_STX1a,_CPLX1,_CPLX3)</td> <td>complex</td> <td>6.0</td> <td>3.0</td> <td>SNAP25, STX1A, VAMP2</td> </tr> <tr> <td>SNARE_complex_(VAMP2,_SNAP25,_STX1a,_CPLX3,_CPLX4)</td> <td>complex</td> <td>6.0</td> <td>3.0</td> <td>SNAP25, STX1A, VAMP2</td> </tr> <tr> <td>SNARE_complex_(VAMP2,_SNAP25,_STX1a,_CPLX1,_CPLX3,_CPLX4)</td> <td>complex</td> <td>8.0</td> <td>3.0</td> <td>SNAP25, STX1A, VAMP2</td> </tr> <tr> <td>SNARE_complex_(VAMP2,_SNAP25,_STX1a,_CPLX1)</td> <td>complex</td> <td>4.0</td> <td>3.0</td> <td>SNAP25, STX1A, VAMP2</td> </tr> <tr> <td>MIR200B</td> <td>miRNA</td> <td>501.0</td> <td>2.0</td> <td>SNAP25, STX1A</td> </tr> <tr> <td>MIR200C</td> <td>miRNA</td> <td>501.0</td> <td>2.0</td> <td>SNAP25, STX1A</td> </tr> <tr> <td>MIR429</td> <td>miRNA</td> <td>501.0</td> <td>2.0</td> <td>SNAP25, STX1A</td> </tr> <tr> <td>MIR520D</td> <td>miRNA</td> <td>667.0</td> <td>2.0</td> <td>SNAP25, VAMP2</td> </tr> </tbody> </table>	Neighbor modules	Type	Module size	Overlap size	Overlap genes	STX1A	hub	59.0	4.0	CPLX2, SNAP25, STX1A, VAMP2	MIR1	miRNA	309.0	3.0	CPLX2, SNAP25, VAMP2	SNARE_complex_(VAMP2,_SNAP25,_STX1a,_CPLX1,_CPLX3)	complex	6.0	3.0	SNAP25, STX1A, VAMP2	SNARE_complex_(VAMP2,_SNAP25,_STX1a,_CPLX3,_CPLX4)	complex	6.0	3.0	SNAP25, STX1A, VAMP2	SNARE_complex_(VAMP2,_SNAP25,_STX1a,_CPLX1,_CPLX3,_CPLX4)	complex	8.0	3.0	SNAP25, STX1A, VAMP2	SNARE_complex_(VAMP2,_SNAP25,_STX1a,_CPLX1)	complex	4.0	3.0	SNAP25, STX1A, VAMP2	MIR200B	miRNA	501.0	2.0	SNAP25, STX1A	MIR200C	miRNA	501.0	2.0	SNAP25, STX1A	MIR429	miRNA	501.0	2.0	SNAP25, STX1A	MIR520D	miRNA	667.0	2.0	SNAP25, VAMP2
Neighbor modules	Type	Module size	Overlap size	Overlap genes																																																				
STX1A	hub	59.0	4.0	CPLX2, SNAP25, STX1A, VAMP2																																																				
MIR1	miRNA	309.0	3.0	CPLX2, SNAP25, VAMP2																																																				
SNARE_complex_(VAMP2,_SNAP25,_STX1a,_CPLX1,_CPLX3)	complex	6.0	3.0	SNAP25, STX1A, VAMP2																																																				
SNARE_complex_(VAMP2,_SNAP25,_STX1a,_CPLX3,_CPLX4)	complex	6.0	3.0	SNAP25, STX1A, VAMP2																																																				
SNARE_complex_(VAMP2,_SNAP25,_STX1a,_CPLX1,_CPLX3,_CPLX4)	complex	8.0	3.0	SNAP25, STX1A, VAMP2																																																				
SNARE_complex_(VAMP2,_SNAP25,_STX1a,_CPLX1)	complex	4.0	3.0	SNAP25, STX1A, VAMP2																																																				
MIR200B	miRNA	501.0	2.0	SNAP25, STX1A																																																				
MIR200C	miRNA	501.0	2.0	SNAP25, STX1A																																																				
MIR429	miRNA	501.0	2.0	SNAP25, STX1A																																																				
MIR520D	miRNA	667.0	2.0	SNAP25, VAMP2																																																				

**Figure 2.** Screenshots of detail information page of SNARE complex. Detail information module includes the description of subunit proteins (A), known gene relationships with their pairwise MI differences (B), graphical view of interaction network (C) and neighboring modules with overlapping genes (D). Disease-associated information of each gene is linked by ‘GAD’ icon (A and B). Green and red colors of gene nodes in (C) represent the degree of up- and downexpressions, respectively. Yellow colored genes in (A) link to Entrez database at NCBI. Yellow colored gene pairs in (B) link to an information page showing gene functions, known gene relationships and associated functional modules.

gene relationships (Figure 2B and C) and neighbor modules that have overlapped genes (Figure 2D), further refined association study among the specific pairs of genes in different functional categories can be designed in detail.

## CONCLUSIONS

DynaMod provides a new method to identify both significant changes of functional modularity by using the average MI differences of gene pairs and differential expressions in predefined functional modules depending on expression profiles under different phenotypes, which

could not be fully supported by typical GSEA approaches. Interestingly, this study showed that the average MI differences acquired by phenotype randomization of an expression profile demonstrate normality in small sampling size. Hence, this proposed method supports that parametric tests such as Z-statistics properly analyze functional modules composed of at least three genes (including at least two biological gene pairs). On the basis of this background, a wide variety of functional modules can be collected and analyzed, including protein complexes, hub-partner groups, miRNA-target groups and transcriptional factor-target groups.

In summary, this tool evaluates the significant modular correlations of various functional categories with gene

expression profiles of different phenotypes using the average MI differences of gene pairs in the modules. Significantly correlated networks across functional modules can be generated through utilization of overlapping genes among different modules. As this tool ascertains the overall effect of those modules, it provides module-wise interpretation of dynamic cellular behaviors. User can conveniently interpret the dynamic modular activities of various functional categories and their networks depending on phenotype changes.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENT

The authors thank anonymous reviews for constructive criticisms and fruitful discussions.

## FUNDING

Grant from the Korea Institute of Science and Technology Information (KISTI). Funding for open access charge: the Korea Institute of Science and Technology Information (KISTI).

*Conflict of interest statement.* None declared.

## REFERENCES

- Hwang,T., Sun,C.H., Yun,T. and Yi,G.S. (2010) FiGS: a filter-based gene selection workbench for microarray data. *BMC Bioinformatics*, **11**, 50.
- Inza,I., Larranaga,P., Blanco,R. and Cerrolaza,A.J. (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif. Intell. Med.*, **31**, 91–103.
- Kim,S.Y. and Volsky,D.J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**, 144.
- Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
- Kim,S.B., Yang,S., Kim,S.K., Kim,S.C., Woo,H.G., Volsky,D.J., Kim,S.Y. and Chu,I.S. (2007) GAzer: gene set analyzer. *Bioinformatics*, **23**, 1697–1699.
- Subramanian,A., Kuehn,H., Gould,J., Tamayo,P. and Mesirov,J.P. (2007) GSEA-P: a desktop application for gene set enrichment analysis. *Bioinformatics*, **23**, 3251–3253.
- Backes,C., Keller,A., Kuentzer,J., Kneissl,B., Comtesse,N., Elnakady,Y.A., Muller,R., Meese,E. and Lenhof,H.P. (2007) GeneTrail—advanced gene set enrichment analysis. *Nucleic Acids Res.*, **35**, W186–W192.
- Boorsma,A., Foat,B.C., Vis,D., Klis,F. and Bussemaker,H.J. (2005) T-profiler: scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Res.*, **33**, W592–W595.
- Mani,K.M., Lefebvre,C., Wang,K., Lim,W.K., Basso,K., Dalla-Favera,R. and Califano,A. (2008) A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Mol. Syst. Biol.*, **4**, 169.
- Taylor,I.W., Linding,R., Warde-Farley,D., Liu,Y., Pesquita,C., Faria,D., Bull,S., Pawson,T., Morris,Q. and Wrana,J.L. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnol.*, **27**, 199–204.
- Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
- Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
- Sun,C.H., Kim,M.S., Han,Y. and Yi,G.S. (2009) COFECO: composite function annotation enriched by protein complex data. *Nucleic Acids Res.*, **37**, W350–W355.
- Han,Y.W., Sun,C.H., Kim,M.S. and Yi,G.S. (2009) Combined Database System for Binary Protein Interaction and Co-complex Association. *2009 International Association of Computer Science and Information Technology—Spring Conference*, iacsit-sc, pp. 538–542.
- Ewing,R.M., Chu,P., Elisma,F., Li,H., Taylor,P., Clime,S., McBroom-Cerajewski,L., Robinson,M.D., O'Connor,L., Li,M. et al. (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.*, **3**, 89.
- Alfarano,C., Andrade,C.E., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechko,B., Boutilier,K., Burgess,E. et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.*, **33**, D418–D424.
- Breitkreutz,B.J., Stark,C., Reguly,T., Boucher,L., Breitkreutz,A., Livstone,M., Oughtred,R., Lackner,D.H., Bahler,J., Wood,V. et al. (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.*, **36**, D637–D640.
- Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
- Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venuopal,A. et al. (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
- Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. et al. (2007) IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
- Ceol,A., Chatr-Aryamontri,A., Licata,L., Peluso,D., Brigandt,L., Perfetto,L., Castagnoli,L. and Cesareni,G. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.
- Pagel,P., Kovac,S., Oesterheld,M., Brauner,B., Dunger-Kaltenbach,I., Frishman,G., Montrone,C., Mark,P., Stumpflen,V., Mewes,H.W. et al. (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics*, **21**, 832–834.
- Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
- Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.
- Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
- Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc., Series B (Methodological)*, **57**, 289–300.
- Sun,L., Hui,A.M., Su,Q., Vortmeyer,A., Kotliarov,Y., Pastorino,S., Passaniti,A., Menon,J., Walling,J., Bailey,R. et al. (2006) Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer cell*, **9**, 287–300.
- Kopp,F., Dahlmann,B. and Kuehn,L. (2001) Reconstitution of hybrid proteasomes from purified PA700-20S complexes and PA28alpha/beta activator: ultrastructure and peptidase activities. *J. Mol. Biol.*, **313**, 465–471.
- Lei,M. (2005) The MCM complex: its role in DNA replication and implications for cancer therapy. *Curr. Cancer Drug Targets*, **5**, 365–380.
- McMahon,H.T., Missler,M., Li,C. and Sudhof,T.C. (1995) Complexins: cytosolic proteins that regulate SNAP receptor function. *Cell*, **83**, 111–119.